应用变异系数识别睡美人文献研究*

■ 唐洁[1,2,3] 曾静静[1,2]

[1] 中国科学院西北生态环境资源研究院 兰州 730000 [2] 中国科学院兰州文献情报中心 兰州 730000

[3] 中国科学院大学 北京 100049

摘 要:[目的/意义]回顾现有的睡美人文献识别方法,梳理不同方法的优缺点,尝试兼顾准确性与易操作性来改进睡美人文献的识别方法。[方法/过程]基于目前发展较为成熟的Bcp指数识别法,借鉴其利用引文曲线"离散程度"进行识别这一核心思想,引入统计学中的"变异系数"概念,将其应用于不同引文曲线类型的区分,从而提出用以识别睡美人文献的PCV指数。[结果/结论]识别结果显示,PCV指数能够较为简单、准确地识别睡美人文献,且该方法对总被引次数具有较低的依赖性。

关键词:睡美人文献 引文曲线 变异系数

分类号:G252

DOI:10.13266/j.issn.0252-3116.2021.06.010

科学文献的生命周期与老化规律是科学传播研究中的重要内容之一。一般而言,文献会在发表后的几年内被其他文献引用,并逐渐达到引文峰值,之后被引次数下降,直至文献不再被引用[1]。然而,有学者发现存在一类文献,其发表初期鲜少被引用,经历一段蛰伏期之后被引量突增。计量学家A. F. J. van Raan[2]将其称为"科学中的睡美人",并提出了沉睡时长、睡眠深度、唤醒强度3项指标对其进行揭示,此后,这类现象逐渐被定量化、规范化进行研究。

"睡美人"现象的本质是其研究内容属于变革性研究或超前性研究[3],识别睡美人文献有助于完善科学评价体系,鼓励创新性研究,也有助于进一步认识科学信息流动机制并发现潜在的创新点。这赋予了睡美人文献重要的研究价值,也使得睡美人文献的识别工作成为图书情报领域的重要研究内容之一。

1 识别方法概述

目前已有多位学者提出了识别睡美人文献的方法,现有的识别方法可以分为3类[4]:①曲线拟合法:通过数学表达式或适当的曲线类型来拟合单篇文献被引次数的年度分布,以此来识别睡美人文献[1,5];②主观指标法:通过设置指标并人为设定阈值来判断一篇文献是否属于睡美人文献[2,6-7];③客观指标法:通过利用指标数值的大小衡量一篇文献可以被看作是睡美人文献的程度,从而消除了主观设置阈值的随意性[8-15]。

前期相关调研结果显示,在进行独立学科领域的睡美人文献识别工作时,得益于操作简便、识别快速,主观指标法的使用率远高于其他方法,但需要人为设定阈值,存在很强的主观性,且容易造成识别的不全面[10]。相对而言,客观指标法和曲线拟合法规避了界定识别标准时的随意性,识别结果更为准确,但计算过程较为复杂。因此,本文尝试兼顾准确性与易操作性对睡美人文献的识别方法进行进一步的探索。

2 方法提出

近年来,睡美人文献的识别方法呈现出由主观指标向客观指标演变的趋势[16],J. Li和F. Y. Ye[17]也指出,识别睡美人文献时应避免人为设定阈值。在这一背景下,本文借鉴Bcp指数[12]的思想,尝试提出新的识别方法。

2.1 Bcp指数

Bcp指数(公式1)的提出经历了对B指数[10]以及SBc指数[11]的完善,是一种发展较为成熟的客观识别

* 本文系国家自然科学基金面上项目"气候变化科学成果集成研究范式及其实现平台研究"(项目编号:41671535)研究成果之一。

作者简介:唐洁(ORCID:0000-0001-7632-3285),硕士研究生;曾静静(ORCID:0000-0002-2236-3924),副研究员,通讯作者,E-mail:zengjj@llas.ac.cn。

收稿日期:2020-09-21 修回日期:2021-01-05 本文起止页码:93-101 本文责任编辑:杜杏叶

方法。其示意如图 1 所示，其中，参考线 l 为“论文发表当年被引次数点”$(0, C_0)$ 到“年度被引次数累积百分比达到 1 的点”$(t_m, 1)$ 的连线。在公式(1)中，$(1-C_0)/t_m$ 代表参考线 l 的斜率，对于任意 $t<t_m$，计算 l_t 与 C_t 的差值，之后将 $t=0$ 到 $t=t_m$ 的差值相加。

$$Bcp = \sum_{t=0}^{t_m} \frac{1-c_0}{t_m} \cdot t + c_0 - c_t \qquad 公式(1)$$

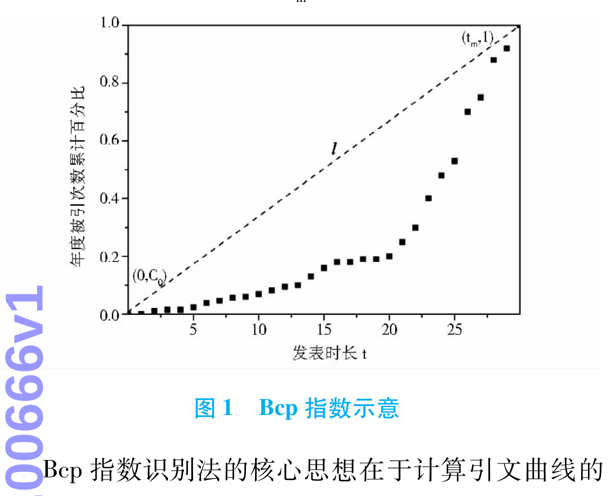


图 1　Bcp 指数示意

Bcp 指数识别法的核心思想在于计算引文曲线的离散程度。睡美人文献具有前期低被引，而后突然高被引的重要特征，这造成其累计引文曲线的离散程度较高[18]，因此，本文引入统计学中的“变异系数”这一概念，以期探索不同引文曲线类型及其离散程度，并依此进行睡美人文献的识别。

2.2　变异系数

变异系数(Coefficient of Variation，CV)又称差异系数或离散系数，是用以表示数据分布离散程度的归一化量度[19]，其定义为各单位的标准差与平均数之比，计算公式为：

$$CV = \frac{\sigma}{\mu} \qquad 公式(2)$$

公式(2)中，σ 为标准差；μ 为平均数。

因变异系数属于无量纲量，其计算过程中消除了不同量纲级别数值的差异，数据之间的可比性较强，这也使得后续工作中构建引文曲线研究框架时，为模拟引文曲线类型而设定的引文量不会对研究结果造成影响，只需着重关注引文曲线的形态。

2.3　不同引文曲线的变异系数

引文曲线又被称为引文模式、引文历史或引文生命周期[1]，通过图形直观地描述引文量随时间的分布变化。已有多位学者对引文曲线类型进行了归纳总结[1,4,20-25](见表 1)。

表 1　引文曲线类型划分

学者	发表时间	划分类型
A. Avramescu	1979	①发表初期即被认可的文献；②被认可程度一般的文献；③几乎未被认可的文献；④发表初期被认可但迅速被摒弃的文献；⑤前期被认可程度低，之后逐步上升的文献[20]
E. S. Aversa	1985	①“缓慢增长－缓慢下降”型；②“快速增长－快速下降”型[21]
V. Cano，N. C. Lind	1991	①早期累积了大部分引用，之后被引量逐渐下降；②被引量在发表初期(前六年)适度增长，之后稳定增长[22]
李江等	2014	①经典型；②指数增长型；③睡美人型；④双峰型；⑤波型[1]
屈文建等	2017	①经典型；②指数下降型；③指数增长型；④睡美人型；⑤多峰型⑥波型[23]
李凌英等	2019	①上升－下降型；②波动型；③下降－上升型；④指数增长型；⑤逐年上升型；⑥延迟认可型[24]
宋呈玉等	2019	①经典型；②指数增长型；③昙花一现型；④睡美人型[4]
熊泽泉	2019	①典型单峰型；②“类睡美人”型；③峰度较低的单峰型；④“缓慢上升－缓慢下降”型(马拉松型)[25]

由表 1 可知，相关研究中分类标准的侧重点有所不同，部分研究以绝对被引量为切入点，根据文章的被认可程度进行引文曲线类型的划分，另一部分研究则侧重于相对被引量进行引文曲线形态的区分。考虑到睡美人文献本身为高被引文献[6,26-29]，并且为了能够较为全面地了解不同类型引文曲线的离散程度，本文对表 1 梳理的引文曲线类型进行归纳后，采取如下研究框架：①经典型：这类文献符合文献老化的一般规律，即发表前期累计大部分引用并达到被引峰值，之后被引量逐渐下降。②昙花一现型：发表后快速到达被引高峰，之后被引次数迅速下降。③指数增长型：发表后年度被引量不断递增。此类文献又被称为天才型论文，较为罕见[24]。④睡美人型：文献发表前期低被引，蛰伏一段时间后被引量剧增，也称延迟认证型。⑤多峰型：其引文历史多次出现峰值，也称波动型。

确定研究框架后，对上述 5 种引文曲线进行模拟。假设 5 篇文献均发表于 2000 年，且截至 2019 年共被引用 300 次，根据各类型引文特点绘制其年度被引曲线(见图 2)及累计被引曲线(见图 3)，并分别计算二者的变异系数，计算结果见表 2。

由表 2 的测试结果可知，就年度被引曲线而言，睡美人型和昙花一现型引文曲线的 CV 值高于其他类型，

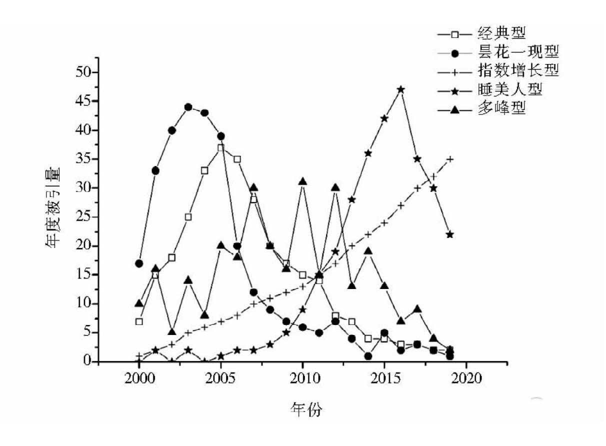


图 2　年度被引曲线

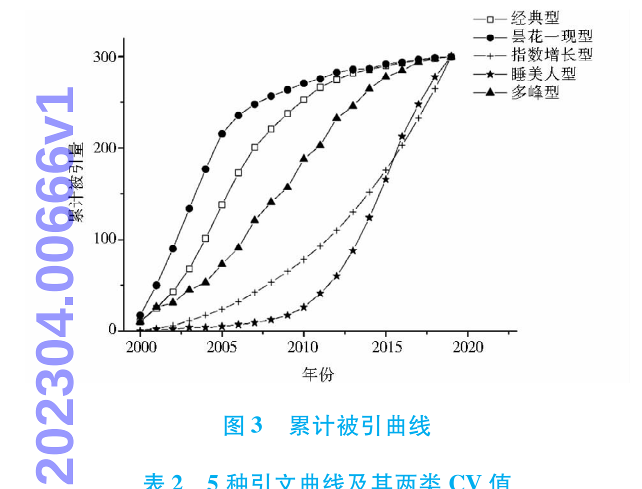


图 3　累计被引曲线

表 2　5 种引文曲线及其两类 CV 值

CV 类型	经典型	昙花一现型	指数增长型	睡美人型	多峰型
年度被引 CV	0.76	1.01	0.68	1.04	0.55
累计被引 CV	0.49	0.37	0.91	1.25	0.61

但昙花一现型文献的累计被引 CV 值却明显低于其他类型。就累计被引曲线而言，睡美人型与指数增长型都具有较高的 CV 值。这一现象表明，仅利用一项引文曲线的 CV 值识别睡美人文献时，有可能会造成其与指数增长型或昙花一现型文献的混淆。为了更加准确地进行睡美人文献的识别工作，需把握其两类 CV 值均较高的这一特性，因此，本文使用两类曲线变异系数的乘积大小来判断一篇文献可以被视作是睡美人文献的程度，并将这一指标记作变异系数之乘积指数，PCV 指数（Product of CV_{yearly} and $CV_{accumulative}$），计算方式如下：

$$PCV = CV_{yearly} \times CV_{accumulative} \quad \text{公式(3)}$$

其中，CVyearly 代表年度被引曲线的变异系数，$CV_{accumulative}$ 代表累计被引曲线的变异系数。

3　实证研究

3.1　选择数据源

本文以 Web of Science（WoS）数据库的核心合集作为数据源，选择 WoS 分类为"information science & library science"，为保证文献有 15 至 25 年的引文窗，限定文献发表时间为 1995 至 2004 年，文献类型为"Article"，共检索出 23 913 篇文献。

由于睡美人文献本身为高被引文献[1,26-29]，这为识别工作前期的数据源筛选提供了启示。本文借鉴普赖斯定律（公式 4）确定高被引文献[30]。

$$N = \sqrt{0.749 \times n_{max}} \quad \text{公式(4)}$$

其中，N 代表高被引文献的最小被引次数，n_{max} 代表文献集中被引量最高论文的被引次数。1995-2004 年，该领域被引最多的文献被引频次为 8 606，计算出 N 为 80.29，最终筛选出 1 098 篇被引频次大于等于 81 的文献。

3.2　识别结果

对获取到的引文数据进行处理并对文献进行编号，编号由文章发表年份后两位数字及该年度文献根据被引次数降序排列的序号组合而成（如文献编号 95-1，代表发表于 1995 年且在 1995 年的文献集中被引频次排名第 1 的文献）。之后进行 PCV 值的计算，1 098篇文献的 PCV 值分布情况如图 4 所示：

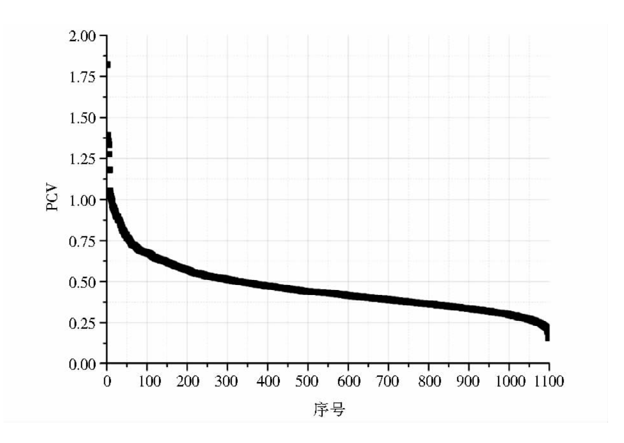


图 4　PCV 值分布情况

由图 4 可知，绝大多数文献的 PCV 数值集中于 0.25至 1.00 区间内，小部分文献的 PCV 值小于 0.25 或大于 1.00。为验证本文提出的识别方法，借鉴已有识别方法相关研究[5]，在目标文献中选取 PCV 值排名 TOP10 的文章进行进一步的识别效果评估，TOP10 文章的相关信息见表 3。

3.3　效果评估

3.3.1　PCV 指数的有效性

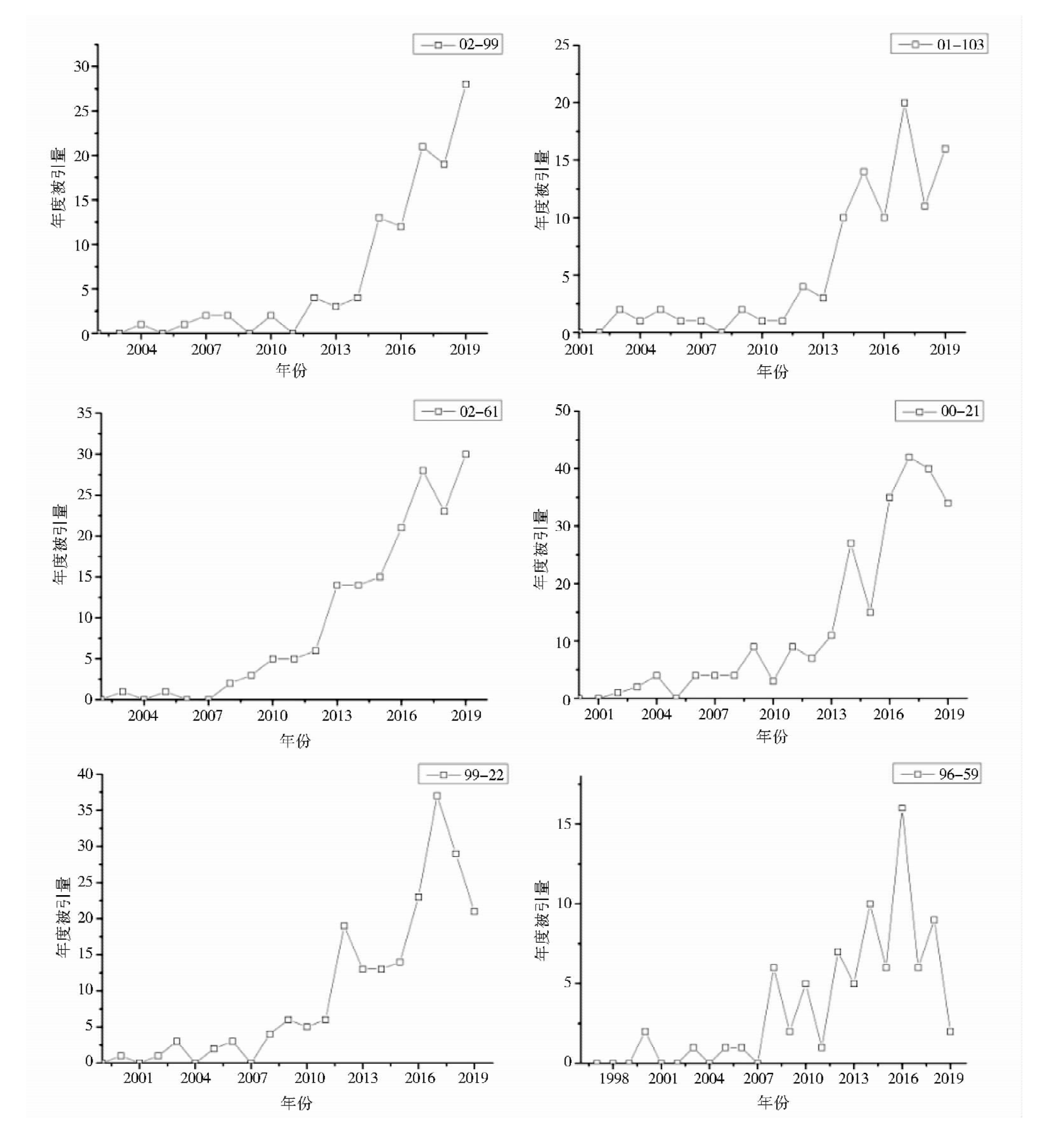
为明确利用 PCV 指数识别出的结果是否符合睡美人文献的特征定义，需要进行有效性检验。目前睡美人文献识别方法的有效性检验主要包括 2 种方式，

表 3 PCV 值排名 TOP10 的文献

序号	PCV 值	文献编号	文章标题
1	1.82	02 – 99	Matrix analysis as a complementary analytic strategy in qualitative inquiry
2	1.39	01 – 103	E-governance and smart communities - A social learning challenge
3	1.35	02 – 61	Searching for safety online：Managing trolling in a feminist forum
4	1.35	00 – 21	Medical subject headings (MeSH)
5	1.33	99 – 22	Knowledge discovery through co-word analysis
6	1.28	96 – 59	A new methodological approach to bibliographic coupling and its application to the national，regional and institutional level
7	1.18	95 – 48	Query evaluation：Strategies and optimizations
8	1.05	01 – 123	Joint and individual interviewing in the context of cancer
9	1.04	99 – 35	Ten tips for reflexive bracketing
10	1.02	01 – 32	Bibliometric cartography of information retrieval research by using co-word analysis

一是进行引文曲线效果分析，二是与其他识别方法进行识别结果重复率的对比。本文将从以上两方面入手检验 PCV 指数法的有效性。

（1）引文曲线效果分析。观察引文曲线的形态能够简单且直观地判断识别方法的有效性，现绘制 PCV 值排名 TOP10 文章的引文曲线（见图 5）。



唐洁，曾静静. 应用变异系数识别睡美人文献研究[J]. 图书情报工作,2021,65(6):93－101.

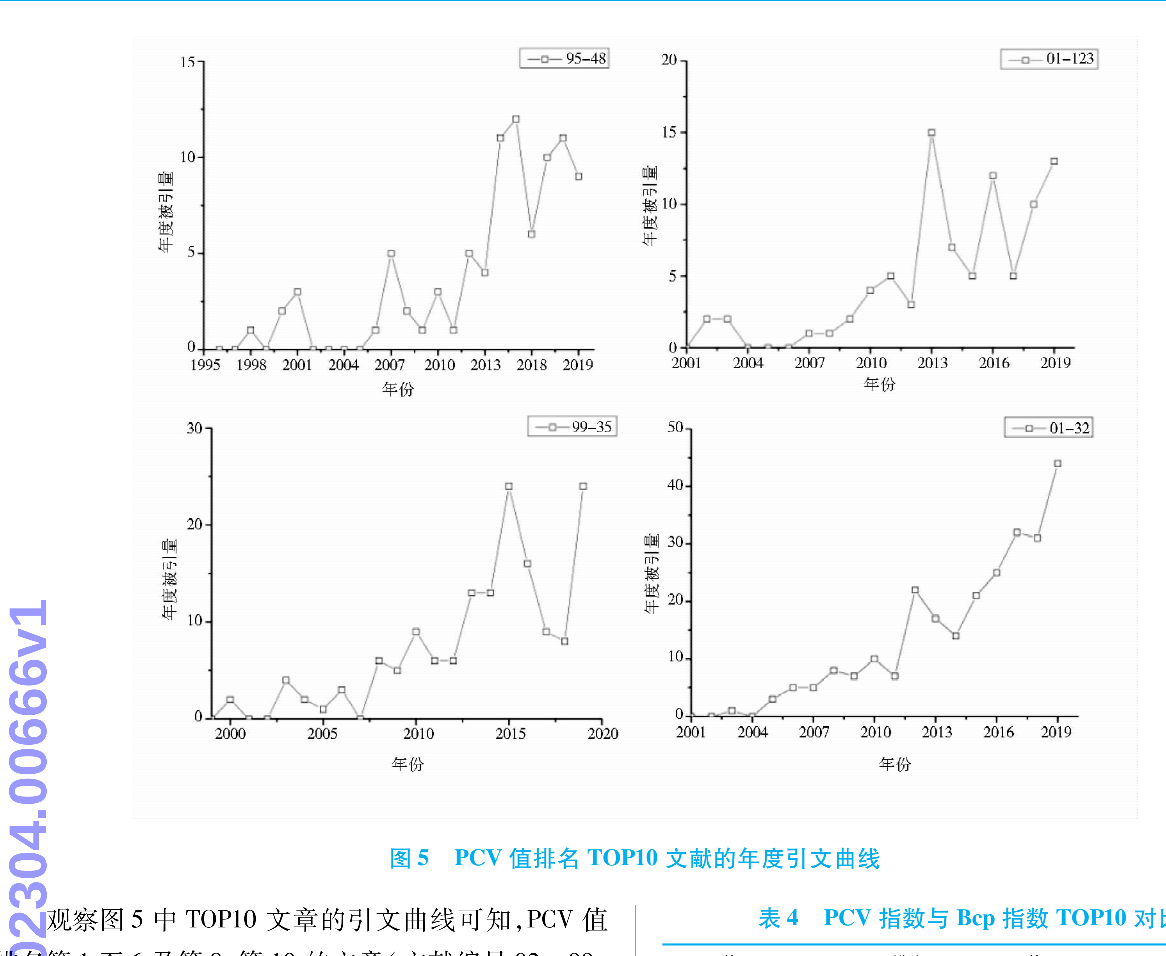


图 5　PCV 值排名 TOP10 文献的年度引文曲线

观察图 5 中 TOP10 文章的引文曲线可知，PCV 值排名第 1 至 6 及第 8、第 10 的文章（文献编号 02－99、01－103、02－61、00－21、99－22、96－59、01－123、01－32）呈现出了较为明显的发表前期低被引、后期突然高被引这一规律，8 篇文献均经历了较为明显的沉睡期。排名为第 7、第 9 的文章（文献编号 95－48、99－35），沉睡期内被引量出现了短暂的波动，但由于增幅较小且持续时间短，其发表前期年均被引量仍然较低。总体而言，TOP10 文章的引文曲线均表现出了睡美人文献的基本特征。

（2）识别结果重复率的对比。考虑到 PCV 指数与 Bcp 指数的识别思路有相似之处，本文将二者识别结果进行了对比。表 4 列出了两种识别框架下排名前 10 的文章及该文章在另一种识别框架下的排名情况。

由表 4 可知，在两种识别框架下排名前 10 的文章中有 5 篇重合，重复率达到 50%。现将已有的不同识别方法识别结果重复率梳理如下表（见表 5）。

由表 5 可知，不同方法识别结果的重复率差异较大，数值最高为 75%，最低为 0%。有学者指出，这种差异性与识别方法的特点以及睡美人引文曲线的形态有关[16]。在本研究中，设置对比范围为 TOP10 的情况下，重复率达到 50%，参考上述研究成果与结论，可以认为本文提出的 PCV 指数识别方法是有效的。

表 4　PCV 指数与 Bcp 指数 TOP10 对比

PCV 值 TOP10	Bcp 排名	Bcp 值 TOP10	PCV 排名
02－99	4	***95－48***	7
01－103	15	***00－21***	4
02－61	18	96－3	17
00－21	2	***02－99***	1
99－22	5	***99－22***	5
96－59	7	95－2	44
95－48	1	***96－59***	6
01－123	51	95－3	46
99－35	25	95－5	40
01－32	27	96－31	21

注：在两种指数 TOP10 中重复出现的文章使用粗斜体标注

3.3.2　PCV 指数与 Bcp 指数的差异性

为进一步探究 PCV 指数与 Bcp 指数在识别睡美人文献方面的差异，本文选取两种计算框架下 TOP10 中相互不重叠的 10 篇文章作为分析对象，借鉴相关研究[2,12]选择如下 6 项指标考察二者的差异性：①发表时长：文章发表年至 2019 年历经的时间跨度（鉴于 2020 年的数据暂不完整，故以 2019 年作为截止时间）。②总被引：文章自发表至 2019 年的总被引次数。③年均被引：总被引量与发表时长的比值。④被引峰值：最高年度被引次数。⑤睡眠时长：借鉴 van Raan[2] 的相关定义，将文章年均被引量处于 0 至 2 次所经历

表 5　各识别方法重复率对比

识别数据	识别方法	对比对象	对比范围(TOP N)	重复率
1970 – 2005 年发表于四大医学名刊的高被引文献[31]	被引速率	B 指数	10	10%
社会科学及商业经济学领域包含"创新"这一关键词的文献[13]	K 指数	3 指标法	53	0%
	K 指数	B 指数	53	25.0%
Science 及 Nature 杂志中的文献[12]	Bcp 指数	B 指数	20	60.0%
WoS 数据库中发表于 1988 – 2007 年的图书情报领域的文献[32]	被引速率	B 指数	35	68.6%
	K 指数	被引速率	35	42.9%
	K 指数	B 指数	35	31.4%
WoS 数据库中发表于 1998 – 2002 年的图书情报领域的文献[5]	曲线拟合法	K 指数	4	75%
WoS 数据库中发表于 1998 – 2003 年的图书情报领域的文献[4]	引文导数法	被引速率	10	60%
WoS 数据库中发表于 1995 – 2004 年的图书情报领域的文献	PCV 指数	Bcp 指数	10	50%

的时长定义为睡眠时长。⑥唤醒强度:文献结束睡眠期后 4 年内的年均被引量。统计以上指标(见表 6)并对其进行独立样本 T 检验,检验结果见表 7。

表 6　排名差异较大的文章指标对比

文章类型	文章编号	PCV 排名	Bcp 排名	发表时长	总被引量	年均被引	被引峰值	睡眠时长	唤醒强度
PCV 排名高于 Bcp 排名	01 – 103	2	15	18	99	5.50	20	14	13.75
	02 – 61	3	18	17	112	6.59	30	10	12.25
	01 – 123	9	51	18	88	4.89	20	14	13.5
	99 – 35	8	25	20	153	7.65	24	10	6.5
	01 – 32	10	27	18	88	4.89	20	14	13.5
Bcp 排名高于 PCV 排名	96 – 3	17	3	23	677	29.43	77	6	8.75
	95 – 2	44	6	24	2009	83.71	167	3	6.25
	95 – 3	46	8	24	1563	65.13	141	2	10.75
	95 – 5	40	9	24	992	41.33	79	8	18.25
	96 – 31	21	21	23	167	7.26	27	3	12.25

表 7　PCV 指数和 Bcp 指数指标差异检验

指标	识别方法	个案数	均值	Sig.(双侧)
发表时长	PCV	5	18.20	0.000(**)
	Bcp	5	23.60	
总被引	PCV	5	108.00	0.040(*)
	Bcp	5	1081.60	
年均被引	PCV	5	5.90	0.042(*)
	Bcp	5	45.37	
被引峰值	PCV	5	22.80	0.039(*)
	Bcp	5	98.20	
睡眠时长	PCV	5	12.40	0.001(**)
	Bcp	5	4.40	
唤醒强度	PCV	5	11.90	0.797
	Bcp	5	11.25	

注:*表示在 0.05 的水平显著;**表示在 0.01 的水平显著

由检验结果可知,PCV 指数与 Bcp 指数识别出的睡美人文献在发表时长、总被引、年均被引、被引峰值 4 项指标上具有显著性差异,Bcp 排名较高的文献以上 4 项指标均高于 PCV 排名较高的文献。这一结果显示,Bcp 指数对于发表时间长且被引量高的文献更敏感,PCV 指数则更易识别出较为年轻的睡美人文献。此外,二者的睡眠时长具有显著性差异,Bcp 排名较高的文献平均睡眠时长为 4.40,而 PCV 排名较高的文献这一数值为 12.40,这说明 PCV 指数识别出的睡美人文献其"沉睡"的特质更为明显。在唤醒强度方面,二者没有显著性差异。

综上分析,PCV 指数法可以成为睡美人文献识别体系的一个有效补充。

4　结论与讨论

本文回顾了现有的睡美人文献识别方法,总结了各类方法的优缺点,为扩充睡美人文献的识别方法体系,提出了 PCV 指数。PCV 指数借鉴 Bcp 指数的识别思想,以衡量引文曲线的离散程度为核心,同时考虑到文献的年度被引曲线和累计被引曲线,进一步降低了识别过程对于文献总被引次数的依赖性,能够更加灵活地发现一些发表年限较短且呈现出"睡美人"特征的文献。同时,在与 Bcp 指数的识别结果进行对比时还发现,PCV 指数识别出的睡美人文献睡眠时间更长。

唐洁，曾静静.应用变异系数识别睡美人文献研究[J].图书情报工作,2021,65(6):93－101.

此外,PCV 指数的计算依托于变异系数,计算简单。综上所述,PCV 指数是一种有效、灵活且易操作的睡美人文献识别方法。

PCV 指数法同样存在一些不足。首先,作为一种客观指标法,PCV 指数法存在与其他客观指标法相同的缺陷,即无法绝对明确地划分睡美人文献与其他类型文献的界限[10]。其次,本文在进行实证研究时选取了高被引文献为数据源,但变异系数本身的特性与效果评估的结果都显示出 PCV 指数对于总被引次数的依赖性极低,因而在识别结果对比中,Bcp 指数识别出了总被引量更高、影响力更大的睡美人文献。针对此问题,未来可以根据具体的研究需求在筛选数据源时进一步提高对于总被引量的限制。

最后,本研究还衍生出了以下问题需要进一步的说明与讨论。

4.1 学科特点对于识别效果的影响

本研究以图情领域的高被引文献为样本,选取样本中的 TOP10 作为分析对象,这 10 篇文章的 PCV 值从 1.02 到 1.82 不等,数值较为分散,且随着数值的降低,文献所呈现出的“睡美人”特性有所减弱,这一现象的产生与选取的研究样本特征有关:一方面,本领域不属于易产生顶级睡美人文献的学科,另一方面,本文限定的引文窗口相对较短,这些因素都有可能影响到最终的识别效果。此外,客观识别法不对文献某一时期的被引量进行数值上的限制,旨在消除主观识别法在设置阈值时的主观性和随意性,但这也造成了睡美人文献与其他文献没有明确的分界线,最终在筛选时不可避免地需要进行人为界定。本文选择 TOP10 为限定标准,但考虑到学科间差异以及人文社科与自然科学领域内睡美人文献的数量差异,这一标准是否适用于其他学科领域仍有待验证。

4.2 引文曲线变异系数的拓展应用

不同类型引文曲线的变异系数计算结果显示,除了睡美人文献之外,还存在其他具有特殊 CV 值的文献类型。例如,昙花一现型文献前期高被引,之后被引量骤降,这使得它同样具有较高的年度被引 CV 值,但由于这类文献较早达到其引文峰值,在此之后因技术更替或研究主题转移等因素很快被遗忘[33],因而后期被引量增长不足,总被引量较早达到稳定状态,最终造成其累计被引曲线的 CV 值低于其他类型的文献,之后可尝试利用上述特点进行相应的识别工作。

4.3 PCV 指数与 Bcp 指数的差异性探究

考虑到 PCV 指数法在识别思路上与 Bcp 指数法具有相似性,本研究进行了二者的识别结果重复率对比,在目标范围内重复率达到了 50%,参考已有重复率对比的结果,该数值处于较高的水平,然而仍略低于预期值,有必要进一步探讨这一现象的成因。从二者的计算过程来看,Bcp 指数为规避对被引次数规模的依赖,将年度被引曲线的纵坐标改为“年度被引次数累积百分比”,但由于其计算包含引文曲线各点到参考线距离的累加过程(见公式(1)),该方法对于发表时间较长的文献更为敏感,这也从某种程度上解释了为何在差异性检验(见表 7)中,Bcp 指数法识别出的文献其发表时长明显更高。然而,PCV 指数由两种引文曲线的变异系数构成,其核心是考察“离散程度”,这也使得 PCV 识别出的部分文献其引文曲线波动较为明显。综上所述,两种方法的思想有相似之处,但其差异性也是客观存在的。

4.4 睡美人文献识别方法的优化

识别睡美人文献的方法体系在不断扩充,然而,在把握睡美人文献基本特征的基础上,不同识别方法的切入点和侧重点有所不同,这造成了识别结果差异的普遍存在。此外,有学者提到,睡美人文献引文曲线的不同阶段受到不同因素的影响,因而其形态多样复杂[17]。综上所述,若将识别方法看作一个睡美人文献的检索系统,做到不误检、不漏检是有一定难度的。为保证识别方法的科学性,客观指标法逐渐替代主观指标法成为主流,但在实践过程中发现,由于其仍然需要人为选择 TOP N 为界限,这也显示出客观指标与主观判断相结合的必要性。此外,近年来多位学者提到通过对现有方法的组合使用来提高识别的准确性与全面性,从而达到互相约束和补充的效果[17],但如何选择识别方法进行组合还需要针对各个研究方法的特点进行实践探究。其次,根据大多数学者在实践中对于客观识别方法的选择倾向,结合穆尔斯定律与齐普夫最省力法则可知,方法的优化不仅在于准确性的提高,其易操作性也不容忽视,未来能否兼顾二者进行睡美人文献的识别亦是必要的尝试。

参考文献:

[1] 李江,姜明利,李玥婷.引文曲线的分析框架研究——以诺贝尔奖得主的引文曲线为例[J].中国图书馆学报,2014,40(2):41－49.

[2] RAAN A F J V. Sleeping beauties in science [J]. Scientometrics,

2004, 59(3):467 –472.

[3] 杜建,孙轶楠,张阳,等. 变革性研究的科学计量学特征与早期识别方法[J]. 中国科学基金,2019,33(1):88 –98.

[4] 宋呈玉,李秀霞,刘黎明. 基于引文曲线导数的睡美人文献识别研究[J]. 情报资料工作,2019,40(3):33 –38.

[5] 宋呈玉,李秀霞,谢瑞霞,等. 基于二次函数曲线拟合的睡美人文献识别研究[J]. 情报杂志,2018,37(6):119 –123 +207.

[6] GARFIELD E. Delayed recognition in scientific discovery: Citation frequency analysis aids the search for case histories[J]. Current contents,1989,12(23):154 –160.

[7] COSTAS R, LEEUWEN T N V, RAAN A F J V. Is scientific literature subject to a 'sell-by-date'? a general methodology to analyze the 'durability' of scientific documents[J]. Journal of the American Society for Information Science and Technology, 2010, 61(2):329 –339.

[8] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3):851 –872.

[9] LI J, SHI D, ZHAO S X, et al. A study of the "heartbeat spectra" for "sleeping beauties" [J]. Journal of informetrics, 2014, 8(3):493 –502.

[10] KE Q, FERRARA E, RADICCHI F, et al. Defining and identifying sleeping beauties in science[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(24): 7426 –7431.

[11] PERUZZO F. Sleeping beauties and the citation dynamics in the network of scientific papers [EB/OL]. [2019 –09 –20] http://tesi. cab. unipd. it/50039/1/Peruzzo_Fabio. pdf.

[12] 杜建,武夷山. 一个用于识别睡美人文献的新的无参数指标——基于"Science"和"Nature"上睡美人文献的验证[J]. 情报理论与实践,2017, 40 (2):19 –25.

[13] TEIXEIRA A A C, VIEIRA P C, ABREU A P. Sleeping beauties and their princes in innovation studies[J]. Scientometrics, 2017, 110(2): 541 –580.

[14] BORNMANN L, YE Y A, YE F Y. Identifying "hot papers" and papers with "delayed recognition" in large-scale datasets by using dynamically normalized citation impact scores[J]. Scientometrics, 2018, 116(2): 655 –674.

[15] YE F Y, BORNMANN L. "Smart girls" versus "sleeping beauties" in the sciences: the identification of instant and delayed recognition by using the citation angle[J]. Journal of the Association of Information Science and Technology, 2018, 69(3): 359 –367.

[16] 宗张建. 睡美人文献识别方法研究进展[J]. 图书情报工作, 2019, 63(16): 132 –142.

[17] LI J, YE F Y. Distinguishing sleeping beauties in science[J]. Scientometrics, 2016, 108(2): 821 –828.

[18] 杜建. 睡美人文献的识别方法与唤醒机制研究[D]. 南京:南京大学, 2017.

[19] 王文森. 变异系数——一个衡量离散程度简单而有用的统计指标[J]. 中国统计, 2007, (6): 41 –42.

[20] AVRAMESCU A. Actuality and obsolescence of scientific literature [J]. Journal of the American Society for Information Science, 1979, 30(5): 296 –303.

[21] AVERSA E S. Citation patterns of highly cited papers and their relationship to literature aging-a study of the working literature[J]. Scientometrics, 1985, 7(3/6): 383 –389.

[22] CANO V, LIND N C. Citation life cycles of ten citation classics [J]. Scientometrics, 1991, 22(2): 297 –312.

[23] 屈文建,胡志伟,周小渝. 面向图情学科热点高被引论文引文曲线特征分析[J]. 情报杂志, 2017, 36(8): 138 –143.

[24] 李凌英,闵超,孙建军. 引文波峰的量化与分布探究[J]. 情报学报, 2019, 38(7): 697 –708.

[25] 熊泽泉,段宇锋. 中文学术期刊论文的引文模式研究——以2006 –2008年图书情报领域期刊论文为例[J]. 图书情报工作, 2019, 63(8): 107 –115.

[26] GLÄNZEL W, SCHLEMMER B. THIJS B. Better late than never? on the chance to become highly cited only beyond the standard bibliometric time horizon[J]. Scientometrics, 2003, 58(3): 571 –586.

[27] ZONG Z J, LIU X Z, FANG H. Sleeping beauties with no prince based on the co-citation criterion[J]. Scientometrics, 2018, 117(3): 1841 –1852.

[28] 杜建,武夷山. 睡美人文献的重要特征、预测线索与政策启示[J]. 科学学研究, 2018, 36(11): 1938 –1945.

[29] 郭斐,鄢小燕. 睡美人文献识别方法分析与改进构想[J]. 图书情报工作, 2016, 60(8): 93 –98.

[30] 钟镇. 从高被引与零被引论文的引文结构差异看 Research Front 与 Research Frontier 的区别[J]. 图书情报工作, 2015, 59(8): 87 –96.

[31] 杜建,武夷山. 睡美人与王子文献的识别方法研究[J]. 图书情报工作, 2015, 59(19): 84 –92.

[32] 李秀霞,邵作运,刘超. 基于K值算法的图书情报领域睡美人文献识别[J]. 图书情报工作, 2017, 61(21): 114 –122.

[33] 李江. 科学中的"睡美人"与"昙花一现"现象评述[J]. 大学图书馆学报, 2016, 34(3): 38 –43.

作者贡献说明：

唐洁：研究设计与规划,数据搜集与分析,撰写论文；
曾静静：研究设计与规划,论文修改与完善。

唐洁，曾静静. 应用变异系数识别睡美人文献研究[J]. 图书情报工作,2021,65(6):93 – 101.

Identify Sleeping Beauties in Science by Coefficient of Variation

Tang Jie[1,2,3] Zeng Jingjing[1,2]

[1] Northwest Institute of Eco-Environment and Resources, CAS, Lanzhou 730000

[2] Lanzhou Information Center, Chinese Academy of Sciences, Lanzhou 730000

[3] University of Chinese Academy of Sciences, Beijing 100049

Abstract: [**Purpose/significance**] This paper aims to review existing identification methods of sleeping beauties in science, discuss strengths and weaknesses of different kinds of methods, and put forward a brand-new method for identifying sleeping papers. [**Method/process**] This study is based on the Bcp index, which is a well-developed and accurate method for identifying sleeping beauties in science. Through referring to the core idea of using the "dispersion degree" of citation curve for identification, the concept of "coefficient of variation" in statistics is introduced to the new method. Then the PCV index is proposed to identify various citation curves, sleeping beauties in particular. [**Result/conclusion**] As is shown in the results, PCV index can effectively identify the sleeping beauties literature. In addition, compared to the Bcp index, the new method has the advantages of simplicity and accuracy, and further reduces the dependence on the total number of citations.

Keywords: sleeping beauty citation curve coefficient of variation